多语种新闻采编系统中跨语言信息检索技术的研究

李满江[1] 任 鹏[2] 张守先[3]
LI Manjiang REN Peng ZHANG Shouxian

摘 要 为满足用户对多语种新闻信息的需求，提高多语种新闻采编系统的信息检索效率，推动跨语言信息检索技术的发展，通过分析跨语言信息检索技术的发展现状和面临的挑战，探讨了跨语言信息检索技术在多语种新闻采编系统中的应用前景和研究价值。跨语言信息检索技术在多语种新闻采编系统中具有广阔的应用前景，深入分析和研究跨语言信息检索技术的关键技术和算法，可以提高新闻采编系统的信息检索效率，满足用户的信息需求，推动跨语言信息检索技术的发展。跨语言信息检索技术在多语种新闻采编系统中具有重要的意义，不断改进和优化这些技术，能够极大地提升新闻采编工作的效率和质量。

关键词 跨语言信息检索；多语种新闻采编系统；语言间相似度计算；机器翻译技术

doi：10.3969/j.issn.1672-9528.2024.08.023

0 引言

随着全球化进程的加速，人类社会逐渐融为一个紧密相连的整体。新闻信息的传播也借此东风，跨越了地域和语言的界限，实现了全球范围内的快速传播。为满足用户对多语种新闻信息的需求，多语种新闻采编系统应运而生。然而，如何在众多语言的新闻信息中快速检索和获取所需内容，成为一个亟待解决的问题。跨语言信息检索（cross-language information retrieval，CLIR）作为一项关键技术，允许系统理解和匹配不同语言的查询和文档，从而为用户提供准确的信息检索服务，它能够帮助用户在不同语言之间查找相关信息，从而打破语言壁垒，提高信息获取的效率。在多语种新闻采编系统中，跨语言信息检索技术具有广泛的应用前景和重要的研究价值。

跨语言信息检索技术主要包括以下几个方面。

（1）语言间相似度计算：通过计算不同语言之间的相似度，衡量文本之间的关联程度[1]。

（2）跨语言检索算法：根据不同语言的文本特点，设计适用于多语言环境的检索算法。

（3）语言无关的索引技术：构建适用于多语言信息的索引结构，提高检索效率。

（4）机器翻译技术：通过自动翻译，将一种语言的文本转换为另一种语言，实现跨语言信息检索。

尽管跨语言信息检索技术已取得了一定的研究成果，但仍面临着许多挑战。语言的复杂性、语义的多样性以及文化背景的差异等因素，都对信息检索的准确性和有效性提出了更高的要求。首先，不同语言之间的语法、词汇和语义差异给跨语言信息检索带来了困难。为了解决这一问题，研究者们提出了基于语法、词汇和语义相似度的计算方法，以提高检索效果[2]。其次，数据稀疏性和语义鸿沟等问题也限制了跨语言信息检索的性能。针对这一挑战，研究者们采用了基于深度学习、聚类分析和特征选择的算法，以提高检索性能。采用基于深度学习的跨语言表示学习算法，以提高不同语言之间的语义相似性计算精度。同时，也可以结合自然语言处理技术，对新闻文本进行更加深入的分析和理解，从而进一步提高检索效果。

跨语言信息检索技术在多语种新闻采编系统中具有广泛的应用前景。通过深入分析和研究跨语言信息检索技术的关键技术和算法，可以为提高新闻采编系统的信息检索效率、满足用户的信息需求、推动跨语言信息检索技术的发展做出贡献。同时，跨语言信息检索技术在实际应用中也面临着诸多挑战，需要不断优化和改进。

1 相关工作

跨语言信息检索是一门交叉学科，其研究始于 20 世纪 90 年代，旨在解决用户使用一种语言查询而目标信息存在

1. 潍坊北大青鸟华光照排有限公司 山东潍坊 261061
2. 潍坊日报社 山东潍坊 261000
3. 半岛都市报社 山东青岛 266071

[基金项目] 山东省重点研发计划山东省科技型中小企业创新能力提升工程项目“智能多语种新闻采编系统”（2022TSGC2368）

于另一种语言环境下的问题。早期的 CLIR 技术主要依赖于词典翻译、统计机器翻译等方法，而近年来，随着深度学习技术的兴起，尤其是神经机器翻译（neural machine translation，NMT）和词嵌入技术（如 Word2Vec，BERT 等）的应用，CLIR 的性能得到了显著提升，能够在保持较高准确率的同时，提高翻译速度和语义理解的深度。在新闻采编领域，多语种处理技术的研究热点包括但不限于自动翻译系统、多语言新闻摘要生成以及基于多模态信息的新闻内容理解。例如，Google 的新闻聚合服务就利用先进的翻译技术，实现了新闻文章的即时多语言版本生成，极大地提升了新闻的国际传播力。此外，一些研究还探索了如何利用社交媒体数据、用户行为分析等方法，来优化新闻内容的多语种推荐策略，从而提高用户的参与度和满意度 [3]。

早期的跨语言信息检索技术的研究主要集中在基于词典的跨语言信息检索方法上，通过构建双语词典来实现不同语言之间的映射。然而，这种方法存在词典规模有限、难以处理语义差异等问题。随着自然语言处理技术的发展，基于机器翻译的跨语言信息检索方法逐渐成为研究热点 [4]。这种方法将查询语句翻译成目标语言，然后在目标语言中进行信息检索，从而实现了跨语言信息检索。基于深度学习的跨语言信息检索方法能够自动学习不同语言之间的语义表示，从而实现了更加准确和高效的跨语言信息检索。这些技术不仅关注词语的直接翻译，还试图理解上下文意义和捕捉不同语言之间的语义关联。此外，还有一些研究关注跨语言信息检索的性能优化和实际应用。例如，通过改进检索算法、优化系统架构等方式来提高跨语言信息检索的准确性和效率。跨语言信息检索技术的相关技术标准和评估方法也在不断完善，为技术的发展和应用提供了重要的参考依据 [5]。

尽管跨语言信息检索技术有了显著的进步，现有的 CLIR 技术仍然面临若干挑战。首先，词义消歧是一个长期存在的问题，特别是在没有足够上下文的情况下确定单词的确切含义。其次，面对不同的语言结构和语法规则，设计一个有效的跨语言匹配模型也是一个挑战。此外，多语种数据集的稀缺性和质量不一也严重制约了 CLIR 系统的性能。构建高质量的平行语料库是 CLIR 技术发展的基石。欧委会的 Europarl、联合国的 UN Parallel Corpus 等大型多语种语料库为 CLIR 模型的训练提供了宝贵资源。然而，对于某些小众语言或低资源语言，语料库的缺乏仍然是一个重大挑战，这要求研究者探索半监督学习、迁移学习等技术来缓解数据不足的问题。为了解决目前遇到的这些问题，研究人员提出了多种方法。其中一些方法聚焦于改进翻译模型，如利用双语对照语料库进行训练来增强模型的语言理解能力 [6]。其他方法则尝试通过引入中间语义表示来克服语言差异，例如使用分布式词嵌入模型来捕获跨语言的语义相似性。还有研究专注于查询扩展技术，通过增加与原始查询相关的词汇来提高检索效果。

跨语言信息检索技术在多语种新闻采编系统中的应用，受到自然语言处理、机器翻译、数据挖掘等多个领域的共同推动。面对多语言环境下的信息检索挑战，未来的研究需继续探索更高效、更智能的解决方案，同时注重技术的普适性和对小众语言的支持，以促进全球新闻传播的无障碍和多元化。

2 跨语言信息检索技术概述

跨语言信息检索是信息检索领域的一个重要分支，它涉及处理和检索不同语言的数据。CLIR 技术使得用户能够以一种语言提交查询，检索出用另一种或多种语言书写的信息的技术。这一技术打破了传统信息检索的语言界限，为用户提供了更为广泛和多元的信息资源。

跨语言信息检索技术的研究涉及了语言学、情报学、计算机科学等多门学科知识，是一个综合性强、富有挑战性的研究领域。其实现主要依赖于信息检索、文字处理、和机器翻译等技术，如文字切分技术、词汇翻译、词频技术、索引技术等 [7]。

CLIR 技术的发展经历了几个主要阶段。

（1）基于词典的方法：早期 CLIR 依赖于双语词典或多语词典，通过查找单词的对应翻译来构建跨语言的索引。这些方法简单直接，但通常无法处理复杂的语义和上下文关系。

（2）基于统计的方法：随后，研究者开发了基于统计的机器翻译技术，如隐含马尔可夫模型（hidden Markov models，HMM）和基于短语的模型。这些方法使用大量的双语语料库来学习词汇和短语的翻译概率，从而提高了翻译的准确性 [8]。

（3）基于语义的方法：为了解决词义消歧和上下文理解的问题，引入了基于语义的技术。这些技术包括利用同义词词典、本体论和语义网络来增强语言间的概念映射。

（4）机器学习和深度学习方法：近年来，随着机器学习和深度学习的兴起，CLIR 技术得到了显著的改进。神经网络机器翻译（neural machine translation，NMT）模型，如循环神经网络（RNN）和 Transformer 架构，能够更好地理解和处理整个句子的上下文意义。此外，预训练语言模型如 BERT（bidirectional encoder representations from transformers）在跨语言信息检索中展现出了卓越的性能。

还有基于知识图谱的方法，借助知识图谱中丰富的语义关系来辅助跨语言信息检索[9]。

这些技术方法各有其特点和优势，在实际应用中往往会根据具体需求和场景进行综合运用或改进创新，以不断提升跨语言信息检索的效果和性能，满足用户在多语种环境下对信息高效获取的需求。

在跨语言信息检索的过程中，涉及以下几个关键步骤。

（1）查询翻译：这是 CLIR 系统的基础，通常采用机器翻译技术，将用户输入的查询语句转换为目标语言。随着神经机器翻译技术的发展，查询翻译的准确性和流畅性有了显著提升[10]。

（2）多语言索引：为了支持跨语言检索，系统需要为每份文档创建多语言版本的索引，这可能涉及文档预处理、分词、词项权重计算等步骤[11]。

（3）语言识别与处理：系统需要能够识别查询语言和文档语言，并进行相应的语言处理，如词干提取、去除停用词等。

（4）结果融合与排序：检索到的跨语言结果需要通过相关性评分进行排序，确保最相关的结果优先展示给用户。这一步骤可能结合翻译质量、文档内容的语义相似度等多种因素。

（5）用户界面与交互：友好的用户界面设计，支持多语言输入和检索结果的多语言展示，提升用户体验。

面临的问题包括以下几点。

（1）翻译质量：机器翻译的准确性直接影响检索效果，特别是对于具有多种含义的词汇和文化特异性表达。

（2）语言资源不均衡：对于一些小语种或低资源语言，缺乏足够的双语对照语料库，影响翻译模型的训练和性能。

（3）多义性处理：同一词语在不同语言和上下文中可能有不同的意义，正确解析查询意图并找到最相关的文档是一大挑战。

（4）评估与优化：跨语言检索系统的性能评估标准和方法仍在发展中，如何有效衡量翻译质量和检索效果是一大难题。

3 部分关键技术

跨语言信息检索（CLIR）涉及以下一些核心技术。

翻译前置（translation lookahead）：指在实际执行查询之前，对可能的查询词条进行预先翻译的过程。这种方法旨在通过提前确定潜在的翻译选项来加快检索过程，并对齐源语言和目标语言之间的词汇差异[12]。

查询翻译（query translation）：它是 CLIR 中最直接的方法之一，涉及将用户的查询从源语言翻译成目标语言。这通常利用机器翻译系统完成，如基于统计的机器翻译模型、基于规则的翻译系统或最近的深度学习翻译模型。

并行语料库（parallel corpora）：在构建有效的查询翻译系统时，通常需要大量的双语对照文本，即所谓的“平行语料库”。这些语料库为统计机器翻译提供了训练数据，使得翻译模型能够学习不同语言之间的对应关系。

跨语言索引（cross-lingual indexing）：跨语言索引是指为目标语言中的每个文档创建索引，同时考虑它们可能在源语言中的表述。一种常见的方法是使用双语索引，其中包含两种语言的关键词。这样，当用源语言查询时，可以直接映射到目标语言的关键词上。

中间语义表示（interlingual representation）：它不依赖于直接的语言翻译，而是尝试创建一个独立于任何特定语言的语义空间，其中源语言和目标语言的文档都可以被映射和比较。这种方法通常涉及使用多语言嵌入或本体论。

机器翻译系统（machine translation systems）：现代 CLIR 系统经常利用先进的机器翻译技术，尤其是基于深度学习的翻译模型，如序列到序列模型和注意力机制。这些技术可以生成非常流畅且准确的翻译，从而改善检索结果的质量。

语境化查询扩展（contextual query expansion）：为了提高检索的相关性，CLIR 系统可能会实施查询扩展，即在原有查询基础上添加相关词汇。这些词汇可以是同义词、近义词或者是从外部知识源中获取的，目的是捕捉更广泛的语义范围。

交互式 CLIR（interactive CLIR）：在某些情况下，用户可以与 CLIR 系统交互，提供反馈以改进查询的翻译和检索结果。这种交互可以通过用户选择相关文档或者修正翻译错误来实现。

这些技术通常相互结合使用，以提高系统的检索效率和精度。CLIR 技术的发展不断推动着全球信息获取的边界，对于打破语言障碍、促进交流具有重要意义。

4 针对多语种新闻采编系统的跨语言信息检索技术改进

为了满足多语种新闻采编系统的跨语言检索要求，从以下几个方面进行探索和优化，以提升检索的准确度、效率及用户体验。

（1）高级翻译技术的集成，例如利用最新的神经网络架构，如 Transformer 及其变种，提升翻译模型的上下文理解和生成质量，减少翻译错误和语义偏差。针对新闻领域特定的词汇、术语和句式，训练专门的翻译模型，提高专业术语

和新闻风格的翻译准确性。结合上下文信息，进行句子级别的翻译，而非简单的单词或短语翻译，以捕捉新闻内容的完整性和准确性。

（2）多模态信息融合检索，结合图像、视频等多模态内容的分析，建立跨语言的语义关联，提升对多媒体新闻内容的检索能力。利用计算机视觉技术提取图片和视频中的关键视觉元素，与文本描述相结合，增强检索的全面性和准确性。

（3）语义增强检索策略，应用自然语言理解技术，如语义分析、实体识别和关系抽取，增强查询的理解深度，自动扩展相关词汇和概念。集成新闻领域知识图谱，利用实体链接和关系推理，提高检索结果的相关性和信息的深度。

（4）用户个性化与交互优化，根据用户的历史查询、阅读偏好和行为模式，动态调整检索策略，提供个性化的新闻推荐。设计智能提示和反馈机制，允许用户在检索过程中修正或细化查询，增强交互体验。

（5）并行语料库与索引优化，持续收集、清洗并标注高质量的多语言新闻语料，为模型训练和评估提供坚实基础。采用倒排索引、布隆过滤器等技术，优化索引结构，加快查询处理速度，支持大规模数据集的快速检索。收集和整理多语种新闻资源，包括新闻文本、图片、视频等，为跨语言信息检索提供丰富的数据支持。建立多语种新闻本体库和术语库，为跨语言新闻采编提供标准化的词汇和术语支持。

（6）构建跨语言知识图谱，将不同语言的新闻实体和概念进行对齐和关联。利用知识图谱中的实体和关系信息，为跨语言信息检索提供额外的语义支持。通过知识图谱的推理能力，实现新闻内容的深度理解和分析。

5 结语

针对多语种新闻采编系统的跨语言信息检索技术有着重要的意义。不断改进和优化这些技术，能够极大地提升新闻采编工作的效率和质量。先进的翻译前置技术与深度学习相结合，为准确地查询翻译奠定了基础，使检索结果更贴合用户需求。持续丰富和完善的并行语料库为跨语言理解提供了有力的支撑。高效的跨语言索引算法确保了快速的信息检索响应。自然语言处理技术的融入进一步增强了检索的准确性和智能化程度。而用户反馈机制的建立则为技术的持续改进提供了方向。

随着技术的不断发展和创新，多语种新闻采编系统中的跨语言信息检索技术将不断完善，更好地服务于全球新闻行业的发展，为人们提供更加全面、准确、及时的多语种新闻资讯，促进信息在不同语言和文化间的流畅传播与共享。

参考文献：

[1] 祝婷 , 胡建成 . 基于关键词聚类的新闻文本相似度计算 [J]. 成都信息工程大学学报 ,2024,39(2):163-169.

[2] 王卫军 , 宁致远 , 董昊 , 等 . 基于语义相似关系的学科交叉主题识别方法 [J]. 情报学报 ,2024,43(1):34-47.

[3] 史明昊 . 基于深度学习的跨模态新闻检索系统的研究与实现 [D]. 北京 : 北京邮电大学 ,2024.

[4] 李翔 . 高朝阳 . 国外机器翻译研究的知识图谱和发展趋势 [J]. 上海翻译 ,2024(2):41-47.

[5] 赵根亮 . 基于深度学习的跨语种语音合成 [D]. 成都 : 电子科技大学 ,2024.

[6] 张磊 . 基于深度学习和词典定义的义原预测研究 [D]. 郑州 : 战略支援部队信息工程大学 ,2020.

[7] 宋鼎新 . 融合句法短语和命名实体的汉英机器翻译研究 [D]. 大连 : 大连理工大学 ,2023.

[8] 崔丹 , 李舒淇 . 基于 AI 算法的自然语言信息提取 - 翻译 - 校对系统设计 [J]. 现代电子技术 ,2024,47(10):111-116.

[9] 雷啸乾 . 基于知识图谱的个性化推荐算法研究与应用 [D]. 南昌 : 南昌大学 , 2024.

[10] 张李义 , 张震云 . 一种新的跨语言商品信息检索方法在图书搜索中的应用 [J]. 现代图书情报技术 ,2010(1):9-14.

[11] 周强伟 , 施水才 , 王洪俊 . 基于预训练模型的受控文本生成研究综述 [J]. 软件导刊 ,2024,23(4):199-207.

[12] 曾剑平 . 翻译可接受性的语言参照系 [J]. 中国科技翻译 , 2024, 37(2): 33-36+12.

【作者简介】

李满江（1970—），男，河北蠡县人，本科，高级工程师，研究方向：中文信息处理。

任鹏（1971—），男，山东高密人，本科，工程师，研究方向：计算机应用。

张守先（1976—），男，山东五莲人，本科，高级工程师，研究方向：大数据应用、计算机应用。

（收稿日期：2024-06-14）